

Cross-layer Optimization for Ultra-reliable and Low-latency Radio Access Networks

Changyang She, Chenyang Yang and Tony Q.S. Quek

Abstract

In this paper, we propose a framework for cross-layer optimization to ensure ultra-high reliability and ultra-low latency in radio access networks, where both transmission delay and queueing delay are considered. With short transmission time, the blocklength of channel codes is finite, and the Shannon Capacity can not be used to characterize the maximal achievable rate with given transmission error probability. With randomly arrived packets, some packets may violate the queueing delay. Moreover, since the queueing delay is shorter than the channel coherence time in typical scenarios, the required transmit power to guarantee the queueing delay and transmission error probability will become unbounded even with spatial diversity. To ensure the required quality-of-service (QoS) with finite transmit power, a proactive packet dropping mechanism is introduced. Then, the overall packet loss probability includes *transmission error probability*, *queueing delay violation probability*, and *packet dropping probability*. We optimize the packet dropping policy, power allocation policy, and bandwidth allocation policy to minimize the transmit power under the QoS constraint. The optimal solution is obtained, which depends on both channel and queue state information. Simulation and numerical results validate our analysis, and show that setting packet loss probabilities equal is a near optimal solution.

Index Terms

Ultra-low latency, ultra-high reliability, cross-layer optimization, radio access networks

I. INTRODUCTION

Supporting ultra-reliable and low-latency communications (URLLC) has become one of the major goals in the fifth generation (5G) cellular networks [2]. Ensuring such a stringent quality-of-service (QoS) enables various applications such as control of exoskeletons for patients, remote driving, free-viewpoint video, and synchronization of suppliers in a smart grid in tactile

This paper was presented in part at the 2016 IEEE Global Communications Conference [1].

Changyang She and Chenyang Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China (email: {cyshe, cyang}@buaa.edu.cn).

T. Q. S. Quek is with Singapore University of Technology and Design and the Institute for Infocomm Research, 8 Somapah Road, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

internet [3], and autonomous vehicles and factory automation in ultra-reliable machine-type-communications (MTC) [4], despite that not all applications of tactile internet and MTC require both ultra-high reliability and ultra-low latency.

Since tactile internet and MTC are primarily applied for emerging mission critical applications, the message such as “touch” and control information is usually conveyed in short packets, and the reliability is reflected by packet loss probability. The traffic supported by URLLC distinguishes from traditional real-time service in QoS requirement and packet size. For human-oriented applications, the requirements on delay and reliability are medium. For example, in the long term evolution (LTE) systems, the maximal queueing delay and queueing delay violation probability for VoIP are respectively 50 ms and 2×10^{-2} in radio access networks, and the minimal packet size is 1500 bytes [5]. For applications target to close-loop control such as vehicle collision avoidance or factory automation, the end-to-end (E2E) or round-trip delay is around 1 ms, the overall packet loss probability is around $10^{-5} \sim 10^{-9}$ [3, 6], and the packet size is 20 bytes or even smaller [2].

LTE systems were designed for human-oriented applications, where the E2E delay includes uplink (UL) and downlink (DL) transmission delay, coding and processing delay, queueing delay, and routing delay in backhaul and core networks [7]. The radio resources are allocated in every transmit time interval (TTI), which is set to be 1 ms [8]. This means that the packets need to wait in the buffer of base station (BS) more than 1 ms before transmission. Therefore, even if other delay components in backhaul and core networks are reduced with new network architectures [9], LTE systems cannot ensure the E2E or round-trip latency of 1 ms.

A. Related Work

While reducing latency in wireless networks is challenging, further ensuring high reliability makes the problem more intricate. To reduce the delay that caused by transmission and signalling [10], a short frame structure was introduced in [11], and the TTI was set identical to the frame duration. To ensure high reliability of transmission with short frame, proper channel coding with finite blocklength is important. Fortunately, the results in [12] indicate that it is possible to guarantee very low transmission error probability with short blocklength channel codes, at the expense of data rate reduction. By using practical coding schemes like Polar codes [13], the delays caused by transmission, signal processing and coding can be reduced.

Exploiting diversity among multiple links has long been used as an effective way to improve the successful transmission probability in wireless channels. To support the high reliability over fading channels, various diversity techniques have been investigated, say frequency diversity and macroscopic diversity in single antenna systems [14, 15] and spatial diversity in multi-antenna systems [16]. Simulation results using practical modulation and coding schemes in [17, 18] show that the required transmit power to ensure given transmission delay and reliability can be rapidly reduced when the number of antennas at a BS increases.

In all these works, only transmission delay and transmission error probability are taken into account in the QoS requirement. In practice, since the packets arrive at the buffer of the BS randomly, there is a queue at the BS. To control the delay and packet loss caused by both queueing and transmission, cross-layer optimization should be considered [1]. Similar to the real-time service such as VoIP, the required queueing performance of URLLC can be modeled as statistical queueing requirement, characterized by the maximal queueing delay and a small delay violation probability. By using effective bandwidth [19] and effective capacity [20] to analyze performance of tactile internet under the statistical queueing requirement, the tradeoff among queueing delay, queueing delay violation probability and throughput was studied in [21], and UL and DL resource allocation was jointly optimized to achieve the E2E delay requirement in [22]. In both works, the Shannon capacity is applied to derive the effective capacity. However, with short transmission delay requirement, channel coding is performed with a finite block of symbols, with which the Shannon capacity is not achievable. In fact, the results obtained by using network calculus in [23] show that if Shannon capacity is used to design resource allocation, the queueing delay and delay violation probability cannot be guaranteed.

Based on the achievable rate of a single antenna system with finite blocklength channel codes derived in [12], queueing delay/length was analyzed in [24, 25]. For applications with medium delay and reliability requirements, the throughput subject to statistical queueing constraints was studied in [24], where the effective capacity was derived by using the achievable rate with finite blocklength channel codes, and an automatic repeat-request (ARQ) mechanism was employed to improve reliability. An energy-efficient packet scheduling policy was optimized in [25] to ensure a strict deadline by assuming packet arrival time and instantaneous channel gains known *a priori*, while the deadline violation probability under the transmit power constraint was not studied.

B. Major Challenges and Our Contributions

Ultra-low latency and ultra-high reliability requirement in radio access network leads to the following challenges in resource allocation optimization.

First, the required queueing delay and transmission delay are shorter than channel coherence time in typical scenarios of URLLC.¹ This results in the following problems. (1) ARQ mechanism can no longer be used to improve reliability. This is because retransmitting a packet in subsequent frames not only introduces extra transmission delay but also can hardly improve the successful transmission probability when the channels in multiple frames stay in deep fading. (2) Time diversity cannot be exploited to enhance reliability, and frequency diversity may not be scalable to the large number of nodes. Moreover, whether spatial diversity can guarantee the reliability is unknown. (3) The studies in [26] show that when the average delay approaches the channel coherence time, the average transmit power could become infinity, because transmitting packets during deep fading leads to unbounded transmit power. Hence, how to ensure both the ultra-low delay and the ultra-high reliability with finite transmit power is unclear.

Second, the blocklength of channel codes is finite. The maximal achievable rate in finite blocklength regime is neither convex nor concave in radio resources such as transmit power and bandwidth [12, 27]. As a result, finding optimal resource allocation policy for URLLC is much more challenging than that for traditional communications, where Shannon capacity is a good approximation of achievable rate and is jointly concave in transmit power and bandwidth.

Third, effective bandwidth is a powerful tool for designing resource allocation to satisfy the statistical queueing requirement of real-time service [19]. Since the distribution of queueing delay is obtained based on large deviation principle, the effective bandwidth can be used when the delay bound is large and the delay violation probability is small [28]. Therefore, using effective bandwidth to satisfy the queueing requirement of URLLC seems problematic.

In this paper, we propose a cross-layer optimization framework for URLLC. While technical challenges in achieving ultra-low E2E/round-trip delay exist at various levels, we only consider transmission delay and queueing delay in radio access networks, and focus on DL transmission. The major contributions of this work are summarized as follows:

- We show that only exploiting spatial diversity cannot ensure the ultra-low latency and ultra-high reliability with finite transmit power over fading channels. To ensure the QoS with

¹In this scenario, effective capacity can no longer be applied.

finite transmit power, we propose a proactive packet dropping mechanism.

- We establish a framework for cross-layer optimization to guarantee the low delay and high reliability, which includes a resource allocation policy and the proactive packet dropping policy depending on both channel and queue state information. By assuming frequency-flat fading channel model, we first optimize the power allocation and packet dropping policies in a single-user scenario, and then extend to the multi-user scenario by further optimizing bandwidth allocation among users. Moreover, how to apply the framework over frequency-selective channel is also discussed.
- We validate that even when the delay bound is extremely short, the upper bound of the complementary cumulative distributed function (CCDF) of queueing delay derived from effective bandwidth still works for Poisson process and Interrupted Poisson Process (IPP), which is more bursty than Poisson process, and Switched Poisson Process (SPP), which is an autocorrelated two-phase Markov Modulated Poisson Process [29].
- We consider the *transmission error probability* with finite blocklength channel coding, the *queueing delay violation probability*, and the *proactive packet dropping probability* in the overall reliability. By simulation and numerical results, we show that setting packet loss probabilities equal is a near optimal solution in terms of minimizing transmit power.

The rest of this paper is organized as follows. Section II describes system model and QoS requirement. Section III shows how to represent queueing delay constraint with effective bandwidth. Section IV introduces the packet dropping policy, and the framework for cross-layer optimization. Section V illustrates how to apply the framework over frequency-selective channel. Simulation and numerical results are provided in Section VI to validate our analysis and to show the optimal solution. Section VII concludes the paper.

II. SYSTEM MODEL AND QOS REQUIREMENT

Consider a time division duplexing cellular system, which consists of a BS with N_t antennas and $K + M$ single-antenna nodes. The nodes are divided into two types. The first type of nodes are K users, which need to upload packets and download packets from the BS. The second type of nodes are M sensors, which only upload packets. In the cases without the need to distinguish between users and sensors, we refer both as nodes. All the notations to be used throughout the paper are summarized in Table I.

TABLE I
SUMMARY OF NOTATIONS

K	number of users	M	number of sensors
T_c	channel coherence time	T_f	duration of one frame
D_{\max}	required delay bound in radio access network	D_{\max}^q	queueing delay bound
ϕ	duration of DL transmission phase	φ	duration of UL transmission phase
ε_k^q	queueing delay violation probability of the k th user	ε_c^q	transmission error probability of the k th user
ε_c^h	proactive packet dropping probability of the k th user	ε_D	overall packet loss probability
N_t	number of antennas at the BS	W_k	bandwidth allocated to the k th user
n_k^s	blocklength of channel coding of the k th user	W_c	coherence bandwidth
$s_k(n)$	achievable rate with finite blocklength of the k th user in the n th frame	$s_k^\infty(n)$	capacity of the k th user in the n th frame
\mathbf{h}_k	channel vector of the k th user	μ_k	average channel gain of the k th user
g_k	normalized instantaneous channel power gain of the k th user	$P_k(n)$	transmit power allocated to the k th user in the n th frame
N_0	single-sided noise spectral density	u	number of bits in one packet
$f_Q^{-1}(x)$	inverse of Q-function	$f_g(x)$	probability density function of the normalized instantaneous channel power gain
\mathcal{A}_k	a set consists of the indices of the nodes that lie in the area of interest w.r.t. the k th user	$a_i(n)$	the number of packets uploaded to the BS from the i th node
$b_k(n)$	number of packets departed from the k th queue in the n th frame	$Q_k(n)$	queue length in terms of number of packets to the k th user in the n th frame
$E_k^B(\theta_k)$	effective bandwidth of the arrival process to the k th user	θ_k	the QoS exponent of the k th user
$P_{D_k}^{\text{UB}}$	upper bound of the queueing delay violation probability of the k th queue	π_l	probability that there are l packets in the queue
λ_k	average packet rate of the k th Poisson process	λ_k^{on}	average packet rate in the “ON” state of the k th IPP
α^{-1}	average duration of “OFF” state of IPP	β^{-1}	average duration of “ON” state of IPP
α_I^{-1}	average duration of the first state of SPP	α_{II}^{-1}	average duration of the second state of SPP
λ_k^I	average packet rate in the first state of the k th SPP	λ_k^{II}	average packet rate in the second state of the k th SPP
ξ_k	ratio of average arrival rate to service rate of the k th queue	γ_k	required SNR of the k th user
η_k	buffer non-empty probability of the k th queue	P_k^{th}	maximal transmit power that can be allocated to the k th user
C^2	variance coefficient		

Time is discretized into frames. Each frame consists of an UL transmission phase and a DL transmission phase. In the UL phase, all nodes upload their messages with short packets to the BS. In the DL phase, the BS processes the received messages from the nodes that lie in the area of interest of each user, and then transmits the relevant messages to the target users. For example, nodes 2, $K + 1$, and $K + 2$ lie in the area of interest with respect to (w.r.t.) user 1, as shown in Fig. 1, and the BS only transmits the messages from nodes 2, $K + 1$, and $K + 2$ to user 1. The scenario that all nodes are users is a special case of our system model with $M = 0$. Since interference causes severe deterioration of QoS, we consider frequency division multiple

access (FDMA) to avoid interference among the users.²

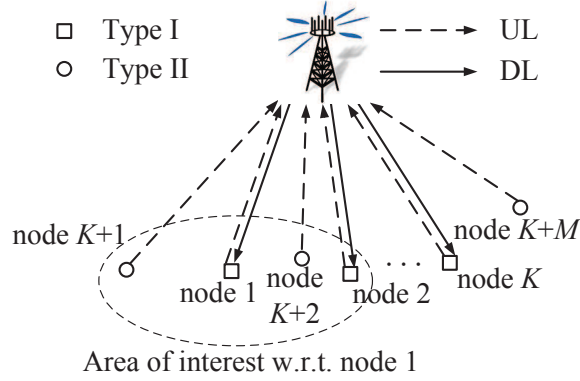


Fig. 1. System model.

The above system model can be applied in analyzing E2E delay in local communication scenarios, where all nodes are associated to adjacent BSs that are connected with each other by fiber backhaul. The delay in fiber backhaul is much less than 1 ms [30], and hence the delay in radio access network dominates the E2E delay. For other communication scenarios (e.g., remote control), the delay components in backhaul and core networks should be taken into account, yet our model can still be used to analyze the delay in radio access [2]. Moreover, the model captures one of the key features of ultra-reliable MTC [4]: a packet generated by one node may be required by multiple users, and one user may also require packets generated by multiple nodes. Therefore, such model is representative for URLLC although cannot cover all application scenarios.³

A. QoS Requirement

The QoS requirement is characterized by the E2E delay for each packet and overall reliability for each user. In the considered radio access network, the E2E delay bound, denoted as D_{\max} , includes UL and DL transmission delay and queueing delay. We only consider one-way delay requirement. By setting D_{\max} less than half of round-trip delay, our studies can be directly extended to the applications with requirement on round-trip delay.

²Although FDMA and time division duplex system is considered in our analysis, it is not hard to extend the framework for cross-layer layer optimization in orthogonal frequency division multiple access and frequency division duplex system, which is widely adopted in LTE-advance radio access networks.

³Direct transmission between nodes (i.e., device-to-device (D2D) communication mode) can help reduce delay with only one hop transmission. However, in D2D mode, the interference becomes more complex than the centralized communication [31]. How to use D2D mode for URLLC deserves further study but is beyond the scope of this work.

Denote the duration of each frame as T_f , the duration of a DL transmission phase as ϕ , and that of an UL transmission phase as φ . To ensure ultra-low transmission delay, we consider the short frame structure proposed in [11], where $T_f \ll D_{\max}$ and the TTI is equal to the frame duration. Owing to the required short delay, retransmission mechanism is unable to be used. The DL transmission of each short packet is finished within the duration of ϕ , and UL transmission of each short packet is completed within φ . If a packet is not successfully transmitted in one frame, then the packet will be lost. Since ϕ and φ are very small, only a few symbols can be transmitted. With finite blocklength channel codes among these symbols, the transmission error is not zero. Since UL transmission policy has been studied in [32], we focus on the DL transmission in this work. Then, the overall reliability for each user, denoted as ε_D , is the overall packet loss probability for each user minus the UL transmission error probability. Denote the DL transmission error probability (i.e. the block error probability [27]) for the k th user as ε_k^c .

Since the transmission delay equals to the frame duration, the queueing delay for every packet should be bounded as $D_{\max}^q \triangleq D_{\max} - T_f$. If the queueing delay bound is not satisfied, then a packet will become useless and has to be dropped. Denote the reactive packet dropping probability due to queueing delay violation as ε_k^q . As detailed later, to satisfy the requirement imposed on the queueing delay for each packet to the k th user, $(D_{\max}^q, \varepsilon_k^q)$, the required transmit power may become unbounded in deep fading. To guarantee the queueing delay with finite transmit power, we proactively select some packets in the queue to discard under deep fading. Denote the proactive packet dropping probability for the k th user as ε_k^h .

Then, the overall reliability for the k th user can be characterized by the overall packet loss probability, which is

$$1 - (1 - \varepsilon_k^c)(1 - \varepsilon_k^q)(1 - \varepsilon_k^h) \approx \varepsilon_k^c + \varepsilon_k^q + \varepsilon_k^h \leq \varepsilon_D, \quad (1)$$

where the approximation is accurate since ε_k^c , ε_k^q , and ε_k^h are extremely small.

B. Channel Model

We consider block fading, where the channel remains constant within a coherence interval and varies independently among intervals. Denote the channel coherence time as T_c . Since the required delay bound D_{\max} is very short, it is reasonable to assume that $T_c > D_{\max} > D_{\max}^q$,

as shown in Fig. 2.⁴ In the following, we consider such a representative scenario for typical applications of URLLC, which is more challenging than the other case with $T_c \leq D_{\max}^q$. Since T_f should be less than D_{\max} and the channel coding is performed within each frame, such a channel (i.e., $T_f < T_c$) is referred to as *quasi-static fading channel* as in [27].

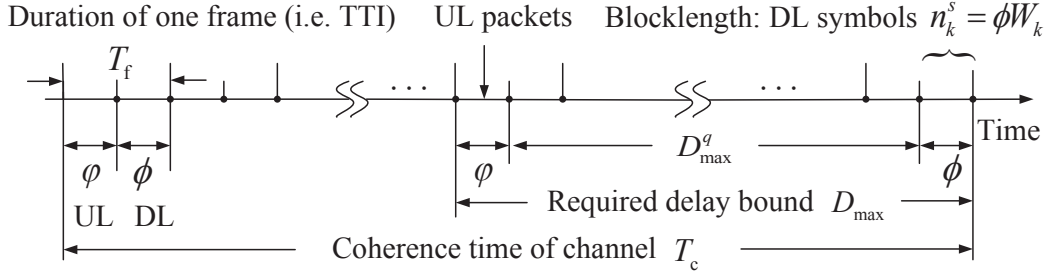


Fig. 2. Relation of the required delay bound, channel coherence time, frame duration, TTI, and blocklength of channel coding.

We consider flat fading channel, which is valid when the bandwidth allocated to each user is smaller than the channel coherence bandwidth. Again, this is a representative scenario for many tactile internet and ultra-reliable MTC applications because the number of users is usually large while the required data rate is low for the purpose of control.

Denote the average channel gain of the k th user as μ_k , and the corresponding channel vector in a certain coherence interval as $\mathbf{h}_k \sim \mathcal{CN}(0, 1) \in \mathbb{C}^{N_t \times 1}$ with independent and identically distributed (i.i.d.) zero mean and unit variance Gaussian elements. Denote the size of each packet as u bits. According to the Shannon capacity formula with infinite blocklength coding, when μ_k and \mathbf{h}_k are perfectly known at the BS, the maximal number of packets that *can be* transmitted to the k th user in the n th frame can be expressed as

$$s_k^\infty(n) = \frac{\phi W_k}{u \ln 2} \ln \left[1 + \frac{\mu_k P_k(n) g_k}{N_0 W_k} \right] \text{ (packets)}, \quad (2)$$

where $P_k(n)$ is the transmit power allocated to the k th user in the n th frame, $g_k = \mathbf{h}_k^H \mathbf{h}_k$, N_0 is the single-sided noise spectral density, and $[\cdot]^H$ denotes the conjugate transpose.

The number of symbols transmitted in DL of one frame (also referred to as the blocklength of channel coding) for the k th user, n_k^s , is determined by the bandwidth and duration, i.e. $n_k^s = \phi W_k$, where W_k is the bandwidth allocated to the k th user. To ensure the ultra-low latency, the transmission duration ϕ is very short. Considering that the bandwidth for each user

⁴For instance, for users with velocities less than 120 km/h in a vehicle communication system operating in carrier frequency of 2 GHz, the channel coherence time is larger than 1 ms, which exceeds the delay bound of each packet. For other applications like smart factory, the velocities of sensors are slow or even zero, and hence $T_c \gg 1$ ms.

is limited, n_k^s is far from infinite, and hence $s_k^\infty(n)$ is not achievable. As shown in [27], the maximal achievable rate with finite blocklength coding is with very complicated expression. By using the *normal approximation* in [27], when n_k^s is finite, the maximal number of packets that *can be* transmitted to the k th user in the n th frame can be accurately approximated as

$$s_k(n) \approx \frac{\phi W_k}{u \ln 2} \left\{ \ln \left[1 + \frac{\mu_k P_k(n) g_k}{N_0 W_k} \right] - \sqrt{\frac{V}{\phi W_k}} f_Q^{-1}(\varepsilon_k^c) \right\} \text{ (packets)}, \quad (3)$$

where $f_Q^{-1}(x)$ is the inverse of Q-function, and V is given by [27]

$$V = 1 - \frac{1}{\left[1 + \frac{\mu_k P_k(n) g_k}{N_0 W_k} \right]^2}. \quad (4)$$

(3) is obtained for interference-free systems, which is valid for the considered FDMA (and also for time division multiple access or space division multiple access with zero-forcing beam-forming). To consider other multiple access techniques where interference cannot be completely avoided, the achievable rate with finite blocklength in interference channels should be used, which however is not available in the literature until now.

As pointed out by the analysis in [23], if (2) is used to design resource allocation, then the queueing delay and the queueing delay violation probability will be underestimated. As a result, the allocated resource is insufficient for ensuring the queueing performance. This indicates that to guarantee ultra-low latency and ultra-high reliability, (3) should be applied.

C. Queueing Model

In the n th frame, the k th user requests the packets uploaded from its nearby nodes. The indices of the nodes that lie in the area of interest w.r.t. the k th user constitute a set \mathcal{A}_k with cardinality $|\mathcal{A}_k|$. As illustrated in Fig. 3, the index set of the nearby nodes of the k th user is $\mathcal{A}_k = \{k+1, \dots, k+m\}$. Then, the number of packets waited in the queue for the k th user at the beginning of the $(n+1)$ th frame can be expressed as

$$Q_k(n+1) = \max \{Q_k(n) - s_k(n), 0\} + \sum_{i \in \mathcal{A}_k} a_i(n), \quad (5)$$

where $a_i(n)$, $i \in \mathcal{A}_k$ is the number of packets uploaded to the BS from the i th nearby node of the k th user.

We consider the scenario that the inter-arrival time between packets could be shorter than D_{\max}^q (otherwise the queueing delay is zero), which happens when the packets for a target user are randomly uploaded from multiple nearby nodes, i.e. $|\mathcal{A}_k| > 1$. At the first glance, such a scenario seems to occur with a low probability. However, to ensure the ultra-high reliability of $\varepsilon_D = 0.001\% \sim 0.00001\%$, the scenario of non-zero queueing delay is not negligible. Denote the number of packets departed from the k th queue in the n th frame as $b_k(n)$. If all the packets in the queue can be completely transmitted in the n th frame, then $b_k(n) = Q_k(n)$. Otherwise, $b_k(n) = s_k(n)$. Hence, we have

$$b_k(n) = \min \{Q_k(n), s_k(n)\}. \quad (6)$$

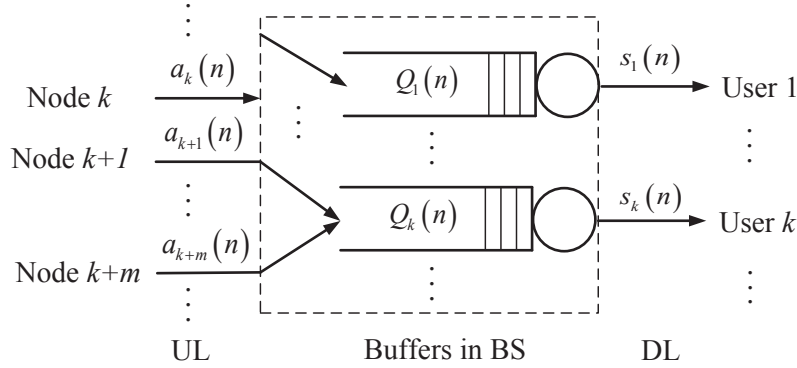


Fig. 3. Queueing model at the BS.

Using (5) and (6), the evolution of the queue length can be described as follows,

$$Q_k(n+1) - Q_k(n) = \sum_{i \in \mathcal{A}_k} a_i(n) - b_k(n), \quad (7)$$

which implies that the queueing delay can be controlled by adjusting the departure process.

III. ENSURING THE QUEUEING DELAY REQUIREMENT

In this section we employ effective bandwidth to represent the queueing delay requirement. We validate that effective bandwidth can be applied in the short delay regime for Poisson arrival process, and then extend the discussion to IPP that is more bursty than the Poisson process, and SPP that is autocorrelated.

A. Representing Queueing Delay Constraint with Effective Bandwidth

For stationary packets arrival process $\{\sum_{i \in \mathcal{A}_k} a_i(n), n = 1, 2, \dots\}$, the effective bandwidth is defined as [19]

$$E_k^B(\theta_k) = \lim_{N \rightarrow \infty} \frac{1}{NT_f\theta_k} \ln \left\{ \mathbb{E} \left[\exp \left(\theta_k \sum_{n=1}^N \sum_{i \in \mathcal{A}_k} a_i(n) \right) \right] \right\} \text{ (packets/s)}, \quad (8)$$

where θ_k is the QoS exponent for the k th user. A larger value of θ_k indicates a smaller queueing delay bound with given queueing delay violation probability.

Remark 1: When the queueing delay bound is not longer than the channel coherence time, the service process is constant within the delay bound with given resources such as transmit power and bandwidth, and the power allocation over fading channel is channel inversion in order to guarantee queueing delay [33]. This is also true when achievable rate in (3) is applied, as explained in what follows. To satisfy the queueing delay requirement of the k th user $(D_{\max}^q, \varepsilon_k^q)$ in fading channels, the constant service rate should be no less than the effective bandwidth of the arrival process of the user. By setting $s_k(n)$ in (3) equal to $E_k^B(\theta_k)$, $P_k(n)g_k$ is constant, i.e., the power allocation is channel inversion, which is not always feasible in practical fading channels. We will show how to handle this issue in the next section.

When the k th user is served with a constant rate equal to $E_k^B(\theta_k)$, the steady state queueing delay violation probability can be approximated as [20]

$$\Pr\{D_k(\infty) > D_{\max}^q\} \approx \eta_k \exp\{-\theta_k E_k^B(\theta_k) D_{\max}^q\}, \quad (9)$$

where η_k is the buffer non-empty probability and the approximation is accurate when $D_{\max}^q \rightarrow \infty$ (i.e. queue length is large enough) [19]. Since $\eta_k \leq 1$, we have

$$\Pr\{D_k(\infty) > D_{\max}^q\} \leq \exp\{-\theta_k E_k^B(\theta_k) D_{\max}^q\} \triangleq P_{D_k}^{\text{UB}}. \quad (10)$$

If the upper bound in (10) satisfies

$$P_{D_k}^{\text{UB}} = \exp\{-\theta_k E_k^B(\theta_k) D_{\max}^q\} = \varepsilon_k^q, \quad (11)$$

then the queueing delay requirement $(D_{\max}^q, \varepsilon_k^q)$ can be satisfied. In other words, if the number of packets transmitted in every frame to the k th user is a constant that satisfies

$$s_k(n) = T_f E_k^B(\theta_k) \text{ (packets)}, \quad (12)$$

then $(D_{\max}^q, \varepsilon_k^q)$ can be ensured [19]. When the k th queue is served by the constant service process $\{s_k(n), n = 1, 2, \dots\}$ that satisfies (12), the departure process in (6) becomes

$$b_k(n) = \min\{Q_k(n), T_f E_k^B(\theta_k)\} \text{ (packets)}. \quad (13)$$

If the departure process $\{b_k(n), n = 1, 2, \dots\}$ satisfies (13), then $(D_{\max}^q, \varepsilon_k^q)$ can be guaranteed. Satisfying (13) does not require constant service process. For example, when $Q_k(n) = 0$, the buffer is empty, then no service is needed.

B. Validating the Upper Bound $P_{D_k}^{\text{UB}}$ in (10) with Representative Arrival Processes

1) *Representative arrival processes:* The aggregation of the periodic packet arrival processes from the $|\mathcal{A}_k|$ users of the k th user (i.e. $\sum_{i \in \mathcal{A}_k} a_i(n)$ in (5)) can be modeled as a Poisson process in vehicle communication and other MTC applications [34, 35]. Denote the average packet rate of the k th Poisson process as λ_k .

Since the features of traffic, say burstiness and autocorrelation, have large impact on the delay performance of queueing systems [36], and the effective bandwidth for real-world arrival processes is hard to obtain, we also consider another two representative traffic models.

As shown in [37], the event-driven packet arrivals in vehicular communication networks can be modelled as a bursty process, IPP. When no event happens, no sensor sends packets to the BS. When an event happens (e.g., a sudden brake) and detected by nearby sensors, the sensors send the packets to the BS. IPP has two states. In the “OFF” state, no packet arrives. In the “ON” state, packets arrive at the buffer of the BS according to a Poisson process with average packet rate λ_k^{on} packets/frame. The durations that the process stays in “OFF” and “ON” states are exponential distributed with mean values of α^{-1} and β^{-1} frames, respectively.

Both Poisson process and IPP are renewal processes, which cannot characterize the autocorrelation of a traffic. In [37], SPP is used to model the aggregation of event-driven packets and periodic packets in vehicle communication networks. Similar to IPP, SPP has two states, where the durations that a SPP stays in the first state and the second state are exponential distributed with mean values of α_I^{-1} and α_{II}^{-1} frames, respectively. In the two states, packets arrive at the buffer of the BS according to Poisson processes with average packet rates λ_k^I and λ_k^{II} packets/frame, respectively. Therefore, a SPP is determined by parameters $(\lambda_k^I, \lambda_k^{II}, \alpha_I, \alpha_{II})$.

The effective bandwidths of Poisson process, IPP and SPP are provided in Appendix A.

Remark 2: Although IPP and SPP come from the vehicle networks, they are also optional models of arrival processes in some tactile internet application scenarios. Haptic communication for remote operation is one of the typical applications in tactile internet. A haptic device that is used for rendering fingertip contact forces consists of several sensors, and force sensor data is acquired at 100 Hz (i.e., 100 packets/sec) [38]. The arrival process depends on the moving patten of one's fingertips. If fingers are either static or moving and switch between these two states, then the arrival process is bursty, and can be modeled as IPP. If the fingers keep moving and the movements of fingers are correlated, then the packets generated by the sensors are also correlated, and can be modeled as SPP.

2) *Validating the upper bound:* The approximation in (9) is accurate when the delay bound is sufficiently large and ε_k^q is very small [19,28]. However, it is unclear how large D_{\max}^q needs to be for an accurate approximation. One possible reason is that it is very difficult to obtain an accurate distribution of the queueing delay.

In fact, what really concerned here is whether the upper bound in (10) is applicable to our problem. If $P_{D_k}^{\text{UB}}$ is indeed an upper bound of $\Pr\{D_k(\infty) > D_{\max}^q\}$, then a transmit policy optimized under the constraint in (12) or (13) can satisfy the queueing delay requirement. In what follows, we derive the queueing delay distribution for Poisson process, which can be used to validate the upper bound in short D_{\max}^q regime numerically.

When a Poisson arrival process is served by a constant service process $\{s_k(n), n = 1, 2, \dots\}$, the well-known M/D/1 queueing model can be applied [39]. For a discrete state M/D/1 queue with integer length (i.e. the number of packets), the closed-form expression of the queue length distribution is known. Specifically, the CCDF of the steady state queue length can be expressed as $\Pr\{Q_k(\infty) > L\} = 1 - \sum_{l=1}^L \pi_l$, where $\pi_l = \Pr\{Q_k(\infty) = l\}$ is the probability that there are l packets in the queue, which is given by

$$\begin{aligned} \pi_0 &= 1 - \xi_k, \quad \pi_1 = (1 - \xi_k)(e^{\xi_k} - 1), \\ \pi_l &= (1 - \xi_k) \left\{ e^{l\xi_k} + \sum_{j=1}^{l-1} e^{j\xi_k} (-1)^{l-j} \left[\frac{(j\xi_k)^{l-j}}{(l-j)!} + \frac{(j\xi_k)^{l-j-1}}{(l-j-1)!} \right] \right\}, \quad (l \geq 2), \end{aligned} \quad (14)$$

with $\xi_k = \lambda_k/s_k(n)$ [39]. For a Poisson arrival process served by a constant service rate $\frac{1}{T_f}s_k(n) = E_k^B(\theta_k)$,

$$\Pr\{D_k(\infty) > D_{\max}^q\} = \Pr\{Q_k(\infty) > E_k^B(\theta_k)D_{\max}^q\}. \quad (15)$$

Then, from (14), the CCDF of the queueing delay can be derived as

$$\Pr\{D_k(\infty) > T_f L / s_k(n)\} = \Pr\{Q_k(\infty) > L\} = 1 - \sum_{l=0}^L \pi_l. \quad (16)$$

The expression of π_l in (14) is too complicated to obtain a closed-form constraint on queueing delay. Nonetheless, (16) can be used to validate the upper bound $P_{D_k}^{\text{UB}}$ in (10) numerically.

Numerical results in [40] show that the upper bound in (10) works for some arrival processes that are more bursty than Poisson process. Later simulations will show that $P_{D_k}^{\text{UB}}$ in (10) can be used to characterize the tail probability of the steady state queueing delay when the packet arrivals are modelled as IPP and SPP.

IV. A FRAMEWORK FOR CROSS-LAYER TRANSMISSION OPTIMIZATION

In this section, we first show that the required transmit power to guarantee the queueing delay and transmission error probability requirement for some packets may become unbounded for any given bandwidth and N_t , owing to $D_{\max}^q < T_c$. To guarantee the QoS in terms of D_{\max}^q and ε_D with finite transmit power, we propose a proactive packet dropping mechanism, and also introduce a power allocation policy depending on both channel information and queue length. Then, we propose a framework to optimize cross-layer transmission strategy, which includes resource allocation and packet dropping policies.

A. Proactive Packet Dropping and Power Allocation

We consider the case where $Q_k(n) \geq T_f E_k^B(\theta_k)$, then $b_k(n) = T_f E_k^B(\theta_k)$. If a transmit power can guarantee such a departure rate, then for the other case where $Q_k(n) < T_f E_k^B(\theta_k)$, $b_k(n) < T_f E_k^B(\theta_k)$ can also be supported, i.e., $(D_{\max}^q, \varepsilon_D^q)$ can be satisfied according to (13).

Substituting $s_k(n)$ in (3) into (12), we can obtain the required SNR γ_k to ensure $(D_{\max}^q, \varepsilon_k^q)$ and ε_k^c for all packets to the k th user using the following equation,

$$\ln(1 + \gamma_k) \approx \frac{T_f u \ln 2}{\phi W_k} E_k^B(\theta_k) + \sqrt{\frac{V}{\phi W_k}} f_Q^{-1}(\varepsilon_k^c). \quad (17)$$

Since $\mathbf{h}_k \sim \mathbb{C}^{N_t}$ is with i.i.d. elements, the channel gain $g_k = \mathbf{h}_k^H \mathbf{h}_k$ follows Wishart distribution [41], whose probability density function is $f_g(x) = \frac{1}{(N_t - 1)!} x^{N_t - 1} e^{-x}$. In the considered typical application scenario with $T_c > D_{\max}^q$, some packets to be transmitted within the delay bound may experience deep fading with channel gain g_k arbitrarily close to zero. Then, the

required transmit power to achieve γ_k in the n th frame, $P_k(n) \triangleq \frac{N_0 W_k \gamma_k}{\mu_k g_k}$, is unbounded. This means that $s_k(n)$ cannot exceed $E_k^B(\theta_k)$ with finite transmit power if the n th frame is in a coherence interval subject to deep fading, even when there is spatial diversity. In other words, for the packets in such an interval, $(D_{\max}^q, \varepsilon_k^q)$ and ε_k^c cannot be guaranteed.

To satisfy the QoS requirement with a finite transmit power, we introduce a proactive packet dropping mechanism.⁵ By “proactive”, we mean that a packet will be intentionally discarded even if its queueing delay has not exceeded D_{\max}^q . By contrast, dropping a packet that has violated the delay requirement is reactive. Then, we can control the total number of packets proactively and reactively dropped to ensure the overall reliability for each user.

Denote the maximal transmit power of the BS as P^{\max} . We discard some packets before transmission in deep fading channels when the required SNR γ_k cannot be achieved with $\sum_{k=1}^K P_k(n) \leq P^{\max}$. However, we can hardly control the packet dropping probability of each user from $\sum_{k=1}^K \frac{N_0 W_k \gamma_k}{\mu_k g_k} \leq P^{\max}$ since the required total transmit power depends on the channel gains of multiple users. To control the packet dropping probability of each user, we introduce the maximal transmit power that can be allocated to the k th user P_k^{th} . When the required transmit power is higher than P_k^{th} , the BS transmits packets to the k th user with power P_k^{th} and drop several packets in the n th frame. Then, the total transmit power of the BS is bounded by $\sum_{k=1}^K P_k^{\text{th}}$.

To ensure $(D_{\max}^q, \varepsilon_k^q)$ and ε_k^c , the *power allocation policy* should depend on both channel gain and queueing length, which is,

$$P_k(n) = \begin{cases} P_k^{\text{th}}, & \text{if } Q_k(n) \geq T_f E_k^B(\theta_k) \text{ and } g_k < \frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}, \\ \frac{N_0 W_k \gamma_k}{\mu_k g_k}, & \text{if } Q_k(n) \geq T_f E_k^B(\theta_k) \text{ and } g_k > \frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}. \end{cases} \quad (18)$$

In the case $Q_k(n) < T_f E_k^B(\theta_k)$, $P_k(n)$ should satisfy $s_k(n) = Q_k(n)$ when $s_k^{\text{th}} > Q_k(n)$ or $P_k(n) = P_k^{\text{th}}$ when $s_k^{\text{th}} \leq Q_k(n)$, where s_k^{th} is the number of packets that can be transmitted in the n th frame with $P_k(n) = P_k^{\text{th}}$. From the approximation in (3), we obtain s_k^{th} as

$$s_k^{\text{th}} \approx \frac{\phi W_k}{u \ln 2} \left\{ \ln \left[1 + \frac{\mu_k P_k^{\text{th}} g_k}{N_0 W_k} \right] - \sqrt{\frac{V}{\phi W_k}} f_Q^{-1}(\varepsilon_k^c) \right\}. \quad (19)$$

⁵The proactive packet dropping is an possible option for ensuring the QoS with finite transmit power in the typical scenario where the required queueing delay bound is shorter than channel coherence time. If we employ a policy that simply does not transmit data when the channel is in deep fading, some packets are lost reactively due to queueing delay violation. How to ensure the queueing delay requirement with extra packets that are lost reactively over deep fading channel is still unknown.

When $g_k < \frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}$ in the n th frame, $s_k^{\text{th}} < T_f E_k^B(\theta_k)$. Since $b_k(n) = \min\{Q_k(n), T_f E_k^B(\theta_k)\}$ needs to be satisfied to ensure $(D_{\max}^q, \varepsilon_k^q)$, the BS has to discard some packets waiting in the queue. Denote the number of packets dropped in the n th frame as $b_k^d(n) = \max\{b_k(n) - s_k^{\text{th}}, 0\}$.

Then, the *proactive packet dropping policy* is

$$b_k^d(n) = \begin{cases} \max(T_f E_k^B(\theta_k) - s_k^{\text{th}}, 0), & \text{if } Q_k(n) \geq T_f E_k^B(\theta_k), \\ \max(Q_k(n) - s_k^{\text{th}}, 0), & \text{if } Q_k(n) < T_f E_k^B(\theta_k). \end{cases} \quad (20)$$

This policy is implemented as follows. If $Q_k(n) \geq T_f E_k^B(\theta_k)$ and $g_k < \frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}$, then P_k^{th} is used to transmit packets and $b_k^d(n)$ packets that cannot be conveyed within the n th frame with P_k^{th} are dropped. With the power allocation policy in (18) and proactive packet dropping policy in (20), the constant service process mentioned in Remark 1 can be provided, which is required to ensure the QoS when the buffer is not empty and the channel gain exceeds a certain threshold. The proactive packet dropping does not lead to extra processing delay since no extra processing is required, and the BS simply drops some packets from the buffer if the channel gain is low.

Despite that the packet dropping probability over large number of frames is very low, the number of packets dropped in a specific coherence interval with deep fading may be large. To avoid extra packet loss over deep fading channel, a possible solution is reducing packets rate at the sources according to the channel gains. However, the BS needs to send CSI to the sources, and this causes extra control overhead.

Similar to the delivery ratio in [42], we define the packet dropping probability as

$$\varepsilon_k^h \triangleq \lim_{N \rightarrow \infty} \frac{\sum_{n=1}^N b_k^d(n)}{\sum_{n=1}^N \sum_{i \in \mathcal{A}_k} a_i(n)} = \frac{\mathbb{E}[b_k^d(n)]}{\mathbb{E}\{\sum_{i \in \mathcal{A}_k} a_i(n)\}}, \quad (21)$$

where the second equality is obtained under the assumption that the queueing system is ergodic, and the average on nominator is taken over both channel gain and queue length.

Based on the analysis in Appendix B, the packet dropping probability can be bounded by

$$\varepsilon_k^h \leq \int_0^{\frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}} \left[1 - \frac{\ln\left(1 + \frac{\mu_k P_k^{\text{th}} g_k}{N_0 W_k}\right)}{\ln(1 + \gamma_k)} \right] f_g(g) dg. \quad (22)$$

B. A Framework for Cross-layer Transmission Optimization

For easy exposition, we first consider single user case, and then extend to multi-user scenario.

With the proactive packet dropping mechanism, the total transmit power is bounded by $\sum_{k=1}^K P_k^{\text{th}}$. To find the minimal resources required to ensure the QoS, we optimize the cross-layer transmission strategy, which includes a transmit power allocation policy $P_k(n)$ and a proactive packet dropping policy $b_k^d(n)$ for single user scenario and also includes a bandwidth allocation policy for multi-user scenario, to minimize $\sum_{k=1}^K P_k^{\text{th}}$ with given total bandwidth of the system.

According to (18), $P_k(n)$ depends on the values of γ_k and P_k^{th} . Given the values of γ_k and ε_k^h , the minimal value of P_k^{th} can be obtained from (22) by letting the equality hold. Moreover, the required SNR γ_k is determined by ε_k^c and ε_k^q according to (17). Therefore, the power allocation policy and the minimal P_k^{th} are uniquely determined by the values of ε_k^c , ε_k^q and ε_k^h .

According to (20), the packet dropping policy depends on s_k^{th} , which can be obtained from (19) and depends on P_k^{th} and ε_k^c . This indicates that to optimize the power allocation policy and packet dropping policy that minimize $\sum_{k=1}^K P_k^{\text{th}}$, we only need to control ε_k^q , ε_k^c , and ε_k^h .

1) *Single-user Scenario*: When $K = 1$, the index k can be omitted for notational simplicity. We consider the case that $Q(n) > 0$. For $Q(n) = 0$, no power is allocated, i.e., $P(n) = 0$.

The values of ε^c , ε^q , and ε^h that minimize P^{th} can be obtained from the following problem,

$$\min_{\varepsilon^q, \varepsilon^c, \varepsilon^h} P^{\text{th}} \quad (23)$$

$$\text{s.t. } \varepsilon^h \leq \int_0^{\frac{N_0 W \gamma}{\alpha P^{\text{th}}}} \left[1 - \frac{\ln \left(1 + \frac{\mu P^{\text{th}} g}{N_0 W} \right)}{\ln(1 + \gamma)} \right] f_g(g) dg, \quad (23a)$$

$$\ln(1 + \gamma) = \frac{T_f u \ln 2}{\phi W} E^B(\theta) + \sqrt{\frac{V}{\phi W}} f_Q^{-1}(\varepsilon^c), \quad (23b)$$

$$\varepsilon^c + \varepsilon^q + \varepsilon^h \leq \varepsilon_D \text{ and } \varepsilon^c, \varepsilon^q, \varepsilon^h \in \mathbb{R}^+, \quad (23c)$$

where constraint (23a) and constraint (23b) are the single-user case of (17) and (22), respectively, $E^B(\theta)$ depends on the source as well as $(D_{\max}^q, \varepsilon^q)$, and \mathbb{R}^+ represents the positive real number.⁶

In the following, we propose a two-step method to find the optimal solution of problem (23).

In the first step, $\varepsilon_0^h \in (0, \varepsilon_D)$ is fixed. Given ε_0^h , P^{th} in the right hand side of (23a) increases with γ . Hence, minimizing P^{th} is equivalent to minimizing γ .

⁶The distribution of channel gain $f_g(g)$ depends on the number of antennas N_t . Therefore, the optimal solution of problem (23) will depend on N_t . We will illustrate the impact of N_t via numerical results in the next section.

For Poisson process, the optimal values of ε^c and ε^q that minimize the required γ can be obtained by solving the following problem,

$$\min_{\varepsilon^q, \varepsilon^c} \frac{T_f u \ln 2 \ln(1/\varepsilon^q)}{\phi W D_{\max}^q \ln \left[1 + \frac{T_f \ln(1/\varepsilon^q)}{D_{\max}^q \lambda} \right]} + \sqrt{\frac{V}{\phi W}} f_Q^{-1}(\varepsilon^c) \quad (24)$$

$$\text{s.t. } \varepsilon^c + \varepsilon^q \leq \varepsilon_D - \varepsilon_0^h, \quad (24a)$$

where the effective bandwidth in (A.2) is used to derive the objective function. As proved in Appendix C, the objective function in (24) is strictly convex in ε^c and ε^q , and hence the problem is convex. To ensure the stringent QoS requirement, the required SNR γ is high, in this case $V \approx 1$ as shown in (4). Then, there is a unique solution of ε^c and ε^q that minimizes γ . Denote the minimal SNR obtained from problem (24) as γ^* . Since the right hand side of (23a) decreases with P^{th} , for given ε_0^h and γ^* , the minimal value of P^{th} can be obtained numerically via binary searching [43] as a function of ε_0^h , denoted as $P^{\text{th}}(\varepsilon_0^h)$.

In the second step, we find the optimal $\varepsilon_0^h \in (0, \varepsilon_D)$ that minimizes $P^{\text{th}}(\varepsilon_0^h)$. Since there is no closed-form expression of $P^{\text{th}}(\varepsilon_0^h)$, exhaustive searching is needed to obtain the optimal ε_0^h in general. However, numerical results indicate that $P^{\text{th}}(\varepsilon_0^h)$ first decreases and then increases with ε_0^h . With this property, we can find the optimal solution of ε_0^h and the required transmit power to ensure ε_D via the exact linear search method [43].

As proved in Appendix D, the solution obtained from the two-step method is the global optimal solution of problem (23) if the solutions of both steps are global optimal.

Impact of traffic feature: To show the impact of burstiness on the cross-layer optimization, we consider IPP with fixed average packet rate in two asymptotic cases, i.e. $C^2 \rightarrow 1$ and $C^2 \rightarrow \infty$, where C^2 is the variance coefficient that can be used to characterize burstiness [29]. To show the impact of burstiness, we keep the average packet rate of IPP, $\frac{\alpha}{\alpha+\beta} \lambda^{\text{on}}$, as a constant. Then, the average packet rate can be expressed as $\frac{\lambda^{\text{on}}}{1+\delta}$, and $C^2 = 1 + \frac{2\delta\lambda^{\text{on}}}{(1+\delta)^2\alpha}$ [29], where $\delta = \beta/\alpha$.

When $\alpha \rightarrow \infty$, $C^2 \rightarrow 1$, the effective bandwidth of the IPP can be expressed as $E^B(\theta) = \frac{\lambda^{\text{on}}}{T_f \theta (1+\delta)} (e^\theta - 1)$, which is the same as the effective bandwidth of a Poisson process with average packet rate $\frac{\lambda^{\text{on}}}{1+\delta}$. When $\alpha \rightarrow 0$, $C^2 \rightarrow \infty$, the effective bandwidth of the IPP can be expressed as $E^B(\theta) = \frac{\lambda^{\text{on}}}{T_f \theta} (e^\theta - 1)$, which is the same as the effective bandwidth of a Poisson process with average packet rate λ^{on} .

To show the impact of autocorrelation, we consider a SPP with parameters $(\lambda^{\text{I}}, \lambda^{\text{II}}, \alpha_{\text{I}}, \alpha_{\text{II}})$,

where $\lambda^I \in [0, \lambda^{\text{on}}]$, $\lambda^{\text{II}} = \lambda^{\text{on}}$, $\alpha_I = \alpha$ and $\alpha_{\text{II}} = \beta$. An upper bound of the effective bandwidth of it can be obtained by substituting $\lambda = \lambda^{\text{on}}$ into (A.1). Therefore, the effective bandwidth of SPP is less than that of a Poisson process with average packet rate $\max\{\lambda^I, \lambda^{\text{II}}\}$.

Remark 3: For IPP, when C^2 increases from 1 to ∞ , the effective bandwidth (i.e. the required constant service rate) increases $1 + \delta$ times. For SPP, the required constant service rate does not exceed the upper bound, which equals to the effective bandwidth of a Poisson process with average packet rate $\max\{\lambda^I, \lambda^{\text{II}}\}$. This indicates that the service rate requirement is still finite for IPP with $C^2 \rightarrow \infty$ or for SPP with any values of α_I and α_{II} . Therefore, the burstiness and autocorrelation will not change the proposed framework.

2) *Multi-user Scenario:* In this case, we jointly optimize W_k , ε_k^c , ε_k^q , and ε_k^h , with which we can obtain the optimal cross-layer strategy including bandwidth allocation, power allocation and packet dropping policies. The optimization problem in the multi-user scenario is formulated as

$$\min_{\substack{W_k, \varepsilon_k^q, \varepsilon_k^c, \varepsilon_k^h \\ k=1,2,\dots,K}} P^{\text{tot}} \triangleq \sum_{k=1}^K P_k^{\text{th}} \quad (25)$$

$$\text{s.t. } \varepsilon_k^h \leq \int_0^{\frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}} \left[1 - \frac{\ln \left(1 + \frac{\mu_k P_k^{\text{th}} g}{N_0 W_k} \right)}{\ln(1 + \gamma_k)} \right] f_g(g) dg, \quad (25a)$$

$$\ln(1 + \gamma_k) = \frac{T_f u \ln 2}{\phi W_k} E_k^B(\theta_k) + \sqrt{\frac{V}{\phi W_k}} f_Q^{-1}(\varepsilon_k^c), \quad (25b)$$

$$\varepsilon_k^c + \varepsilon_k^q + \varepsilon_k^h \leq \varepsilon_D \text{ and } \varepsilon_k^c, \varepsilon_k^q, \varepsilon_k^h \in \mathbb{R}^+, \quad (25c)$$

$$\sum_k^K W_k \leq W^{\text{max}}, W_k \geq W_0, k = 1, \dots, K, \quad (25d)$$

where $W_0 = 1/\phi$ is the bandwidth required to transmit one symbol in a DL phase, (25d) is the bandwidth constraint. Since the number of symbols $n_k^s = \phi W_k$ is an integer, W_k should be divisible by W_0 . As a result, this is a mixed-integer programming problem.

Given the values of the discrete variables $W_k, k = 1, \dots, K$, the problem can be decomposed into K single-user problems similar to (23), which can be solved by the two-step method. Then, the power allocation policy among subsequent TTIs and the packet dropping policy can be obtained similarly to those in the single-user scenario, i.e., (18) and (20). We refer to the K single-user problems as *subproblem I*. Since binary search and exact linear search methods are

applied in solving subproblem I, the complexity of the two-step method is $O(\log_2(\frac{\varepsilon_D}{\Delta^h}) \log_2(\frac{\varepsilon_D}{\Delta^c}))$.⁷

The complexity of problem (25) is determined by the integer programming that optimizes $W_k, k = 1, \dots, K$ with given $\varepsilon_k^c, \varepsilon_k^q, \varepsilon_k^h$ to minimize the objective function in (25). We refer this integer programming as *subproblem II*. Since $W_k \geq W_0$, at least W_0 is allocated to each of the K users. The remaining bandwidth is $W^{\max} - KW_0$. To solve problem (25), we need to allocate the remaining bandwidth to K users. Thus, subproblem II includes around $K^{W^{\max}/W_0 - K}$ feasible solutions. To reduce complexity, a heuristic algorithm is proposed, as listed in Table II. The basic idea is similar to the steepest descent method [43]. The bandwidth allocation algorithm includes $W^{\max}/W_0 - K$ steps. In each step, W_0 bandwidth is allocated to one of the K users that leads to the steepest total transmit power descent. The proposed algorithm only needs to solve subproblem I for $K(W^{\max}/W_0 - K)$ times, and hence the complexity is $O(K(W^{\max}/W_0 - K))$. Further considering the complexity of the two-step method for solving subproblem I, the overall complexity of the proposed algorithm is $O(K(W^{\max}/W_0 - K) \log_2(\frac{\varepsilon_D}{\Delta^h}) \log_2(\frac{\varepsilon_D}{\Delta^c}))$.

TABLE II
BANDWIDTH ALLOCATION ALGORITHM

Input:	Number of users K , total bandwidth W^{\max} , duration of each DL phase ϕ , packet size u , noise spectral density N_0 , number of transmit antennas N_t , average channel gains of users $\mu_k, k = 1, \dots, K$.
Output:	Bandwidth allocation $W_k^*, k = 1, \dots, K$.
1:	Set $n_k^s(0) := 1, k = 1, \dots, K$. Set $l := 1$. Set $W_0 := 1/\phi$.
2:	Solve subproblem I with $W_k(0) = n_k^s(0)W_0$, and obtain the total transmit power $P^{\text{tot}}(0)$.
3:	while $l \leq W^{\max}/W_k(0) - K$ do
4:	Set $\hat{k} := 1$
5:	while $\hat{k} \leq K$ do
6:	$n_{\hat{k}}^s(l) := n_{\hat{k}}^s(l-1) + 1; n_k^s(l) := n_k^s(l-1), k \neq \hat{k}$.
7:	Solve subproblem I with $W_k(l) = n_k^s(l)W_0$, and obtain $\hat{P}_{\hat{k}}^{\text{tot}}(l)$.
8:	$\hat{k} := \hat{k} + 1$.
9:	end while
10:	$k^* := \arg \min_{\hat{k}} \hat{P}_{\hat{k}}^{\text{tot}}(l)$.
11:	$n_{k^*}^s(l) := n_{k^*}^s(l-1) + 1; n_k^s(l) := n_k^s(l-1), k \neq k^*$.
12:	$l := l + 1$.
13:	end while
14:	return $W_k^* = n_k^s(l-1)W_0, k = 1, \dots, K$.

V. APPLYING THE FRAMEWORK OVER FREQUENCY-SELECTIVE CHANNEL

The bandwidth allocation depends on the number of users and the average channel gains of each user. For a user located at the edge of a cell, the bandwidth allocated to it (i.e., W_k in

⁷The complexity of a searching algorithm depends on the stopping criterion. Here, the iterations stop if $|\varepsilon_k^h(i) - \varepsilon_k^h(i+1)| < \Delta^h$ or $|\varepsilon_k^c(i) - \varepsilon_k^c(i+1)| < \Delta^c$ is satisfied, where $\varepsilon_k^h(i)$ and $\varepsilon_k^c(i)$ are the results obtained after i iterations.

problem (25)) could be larger than the coherence bandwidth if the number of users is not very large. In this section, we show how to apply the framework over frequency-selective channel.

Denote the number of subchannels allocated to the k th user as N_k^c . The bandwidth of each subchannel is W_c (i.e., coherence bandwidth). To study the delay and reliability performance, we need first find the achievable rate with finite blocklength. As shown in Appendix E, the number of packet that can be transmitted in one frame can be obtained based on the results in [27], i.e.,

$$s_k^{\text{fs}} \approx \frac{\phi W_c}{u \ln 2} \left\{ \sum_{j=1}^{N_k^c} \ln \left[1 + \frac{\mu_k P_{kj}(n) g_{kj}}{N_0 W_c} \right] - \sqrt{\frac{V}{\phi W_c}} f_Q^{-1}(\varepsilon_k^c) \right\} \text{ (packets)}, \quad (26)$$

where $P_{kj}(n)$ is the transmit power allocated on the j th subchannel of the k th user in the n th frame, g_{kj} is the instantaneous channel gain on the j th subchannel of the k th user, and $V = N_k^c - \sum_{j=1}^{N_k^c} \frac{1}{\left[1 + \frac{\mu_k P_{kj}(n) g_{kj}}{N_0 W_c} \right]^2}$. Since the channel gains on all the subchannels could be arbitrarily close to zero, the required transmit power to guarantee queueing delay is also unbounded.

To transmit s_k^{fs} packets in one frame, the channel coding on different subchannels are correlated. In other words, if the packets transmitted over one subchannel is not decoded successfully, the packets transmitted over the other subchannels are lost. To avoid that all the packets transmitted in one frame are missing at the same time, we consider the systems where the channel coding on each subchannels is independent of the others. When the number of packets transmitted over each subchannel is $E_k^B(\theta_k)/N_k^c$, the constraints on proactive packet dropping probability, queueing delay violation probability and transmission error probability can be obtained by replacing W_k and $E_k^B(\theta_k)$ in (25a) and (25b) with W_c and $E_k^B(\theta_k)/N_k^c$. In this way, the proposed framework can be further applied over frequency-selective channel.

VI. SIMULATION AND NUMERICAL RESULTS

In this section, we first validate that the effective bandwidth can be used as a tool to optimize resource allocation in short delay regime for Poisson process, IPP and SPP. Then, we show the optimal values of ε_k^q , ε_k^c and ε_k^h , and the required maximal transmit power for both Poisson process and IPP.⁸ Next, we compare the required transmit power of the proposed algorithm with the global optimal policy obtained by exhaustive searching.

⁸The optimal values of ε_k^q , ε_k^c and ε_k^h and the required transmit power for SPP are similar to that for IPP, and hence the results for SPP are omitted for conciseness.

A single-BS scenario is considered in this section. The users are uniformly distributed with distances from the BS as 50 m \sim 200 m. The arrival process of each user is modeled as Poisson process, IPP, or SPP with average rate 1000 packets/s. Such an average rate is representative in two typical application scenarios. The first one is the vehicle safety, where each user requests the safety messages from 50 nearby sensors, and each sensor uploads packets to the BS with average rate 20 packets/s [37]. The other one is haptic communication, where each user requests the packets from 10 nearby sensors, and each sensor acquires fingertip contact forces at 100 Hz (i.e., 100 packets/s) [38]. Other parameters are listed in Table III, unless otherwise specified.

TABLE III
PARAMETERS [6, 37]

Overall reliability requirement ε_D	1 – 99.99999%
E2E delay requirement D_{\max}	1 ms
Queueing delay requirement D_{\max}^q	0.9 ms
Duration of each frame (equals to TTI)	0.1 ms
Duration of downlink phase	0.05 ms
Single-sided noise spectral density N_0	–173 dBm/Hz
Packet size u	20 bytes
Path loss model $10 \lg(\mu_k)$	$35.3 + 37.6 \lg(d_k)$
Average duration of “OFF” state α^{-1}	1 s (i.e. 10^4 frames)
Average duration of “ON” state β^{-1}	1 s (i.e. 10^4 frames)

The CCDFs of queue length and queueing delay for the packets to the k th user are shown in Fig. 4, where (15) is used to translate the CCDF of the queueing delay into the CCDF of queue length. To obtain the upper bound $\Pr\{D_k(\infty) > D_{\text{th}}\} \leq \exp\{-\theta_k E_k^B(\theta_k) D_{\text{th}}\}$ is computed by changing D_{th} from 0 to D_{\max}^q . The CCDFs of queueing delay are obtained via Monte Carlo simulation by generating arrival process and service process during 10^{10} frames. Numerical results in Fig. 4(a) indicate that for Poisson process, the upper bound derived by effective bandwidth works when the maximal queue length is short. Simulation results in Fig. 4(b) show that the upper bound also works for IPP and SPP. In fact, it has been observed in [44] that effective bandwidth can be used to design resource allocation policy under statistical queueing delay requirement when D_{\max}^q is small, if the TTI is much shorter than the delay bound.

The optimal solution of problem (23) and the required maximal transmit power for both Poisson and IPP are shown in Fig. 5. The results in Fig. 5(a) show that ε_k^c , ε_k^q and ε_k^h are in

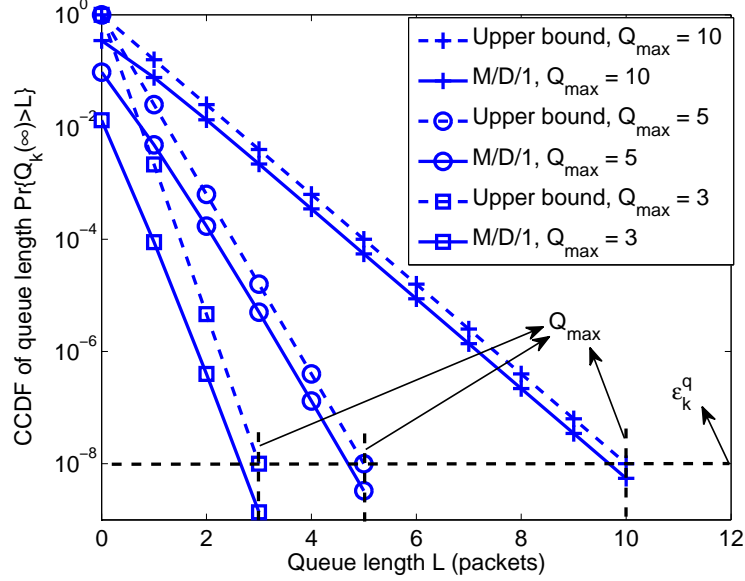
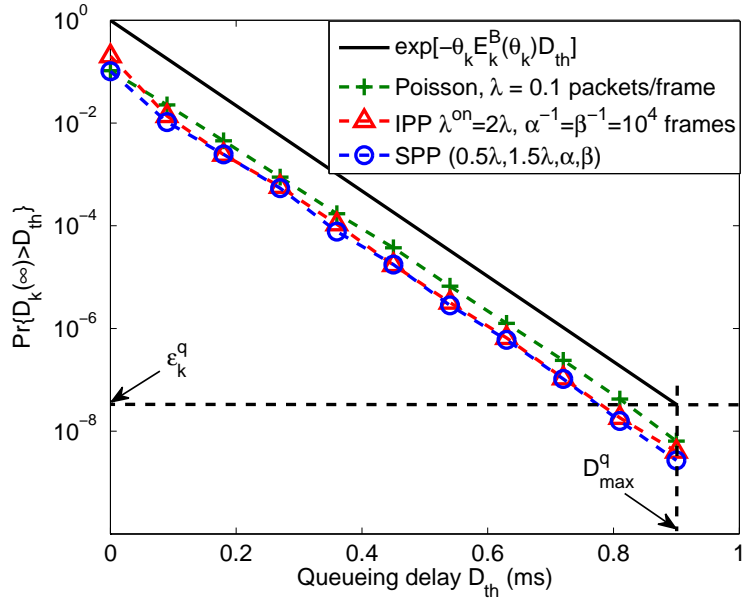
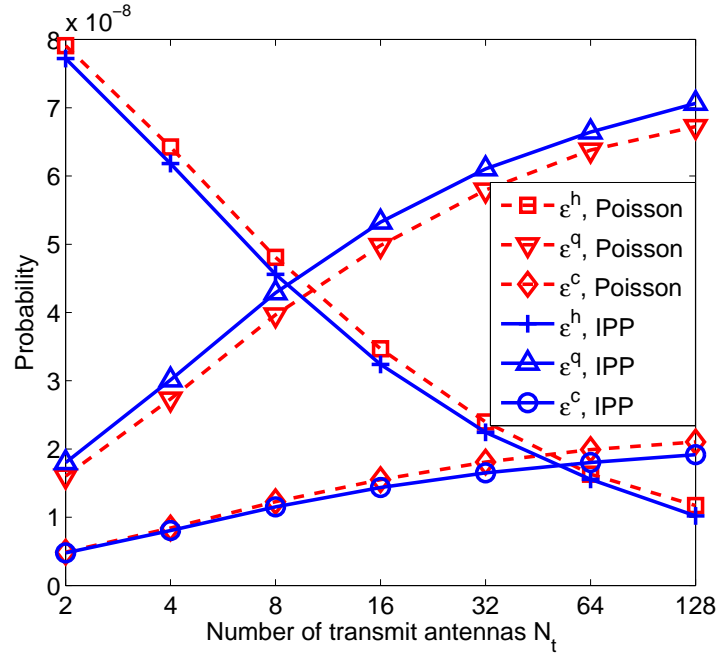
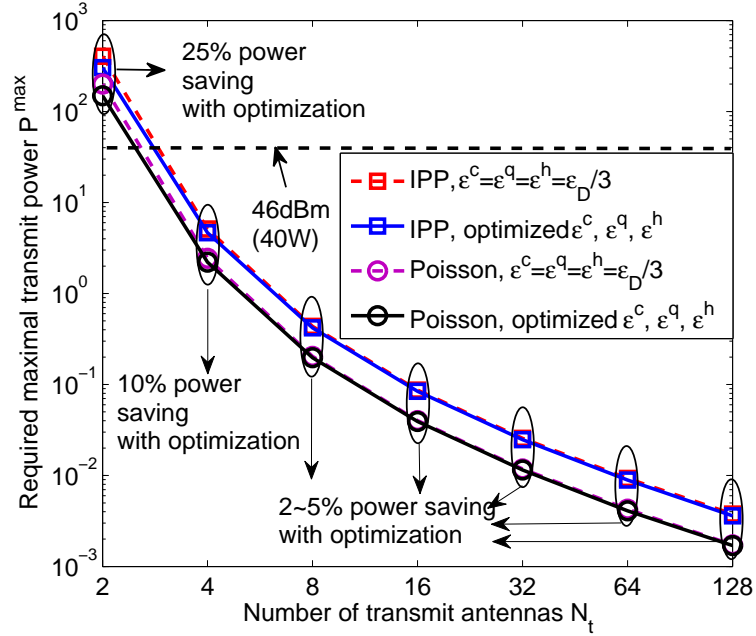
(a) Poisson arrivals, where $\varepsilon_k^q = 10^{-8}$.(b) Poisson arrival, IPP and SPP, where $C^2 = 1001$ for the IPP.

Fig. 4. Validating the upper bound in (10).

the same order of magnitude with different values of N_t . In fact, similar to ε_k^h , when either ε_k^c or ε_k^q is set as zero, the required transmit power will become infinite, because $E_k^B(\theta_k) \rightarrow \infty$ when $\varepsilon_k^q = 0$ (as can be clearly seen from (A.2)) and $f_Q^{-1}(x) \rightarrow \infty$ (and hence $s_k(n)$ in (3) approaches infinity) when $\varepsilon_k^c = 0$. This implies that the optimal probabilities will also be in the same order when other system parameters than N_t change. On the other hand, Fig. 5(b) shows



(a) Optimal values of ε_k^c , ε_k^q and ε_k^h that minimize the required transmit power.



(b) Required maximal transmit power.

Fig. 5. Single-user scenario, where user-BS distance is 200 m, bandwidth is 0.5 MHz and $\alpha = \beta$.

TABLE IV
REQUIRED TRANSMIT POWER, $W^{\max} = 1$ MHz, $N_t = 2$

Number of users K	2	4	6
Proposed Algorithm	0.0216 W	0.155 W	5.26 W
Exhaustive Searching	0.0216 W	0.155 W	5.26 W

that compared with $\varepsilon_k^c = \varepsilon_k^q = \varepsilon_k^h$, the required maximal transmit power only reduces $2 \sim 5\%$ with the optimized ε_k^c , ε_k^q and ε_k^h when $N_t \geq 8$. This implies that dividing the required packet loss probability equally to the three probabilities will cause minor performance loss.

Moreover, the optimal queueing delay violation probability for IPP is higher than that for Poisson process. This indicates that bursty arrival processes lead to higher queueing delay violation probability. Furthermore, P^{th} decreases extremely fast as N_t increases. This agrees with the intuition: increasing the number of transmit antennas is an efficient way to reduce the required maximal transmit power thanks to the spatial diversity.

The required $\sum_{k=1}^K P_k^{\text{th}}$ obtained by the proposed algorithm and the global optimal solution with exhaustive searching are provided in Table IV. The results illustrate that the proposed algorithm is near-optimal. Because the complexity of exhaustive search method is extremely high with large W^{\max} , we only provide results with small values of W^{\max} and K .

VII. CONCLUSIONS

In this paper, we studied how to optimize resource allocation to guarantee the ultra-low latency and ultra-high reliability for radio access networks in typical application scenarios where the required delay is shorter than channel coherence time. Both queueing delay and transmission delay were considered in the latency, and the transmission error probability, queueing delay violation probability, and packet dropping probability were taken into account in the reliability. We first showed that the transmit power is unbounded when queueing delay bound is shorter than channel coherence time. To satisfy the QoS requirement with finite transmit power, a proactive packet dropping mechanism was proposed. A framework for optimizing resource allocation to ensure the stringent QoS was established, where a queue state and channel state information dependent transmit power allocation and packet dropping policies were optimized for single user case, and bandwidth allocation was further optimized for multi-user scenario, to minimize the required maximal transmit power of the BS. How to apply the proposed framework over frequency-selective channel is also illustrated. Simulation results validated that effective bandwidth can be

used to optimize resource allocation for Poisson process, IPP and SPP, which are representative traffic models to characterizing performance of a system with queueing. Numerical results showed that the transmission error probability, queueing delay violation probability, and packet dropping probability are in the same order of magnitude, and a near optimal solution is setting the three packet loss probabilities equal.

APPENDIX A

EFFECTIVE BANDWIDTH OF SEVERAL RELEVANT ARRIVAL PROCESSES

Poisson arrival process: The effective bandwidth of Poisson process is given by

$$E_k^B(\theta_k) = \frac{\lambda_k}{T_f \theta_k} (e^{\theta_k} - 1) \text{ (packets/s)}. \quad (\text{A.1})$$

Substituting (A.1) into (11), we can obtain the required QoS exponent $\theta_k = \ln \left[\frac{T_f \ln(1/\varepsilon_k^q)}{\lambda_k D_{\max}^q} + 1 \right]$. Then, (A.1) can be re-expressed as a function of $(D_{\max}^q, \varepsilon_k^q)$ as

$$E_k^B(\theta_k) = \frac{\ln(1/\varepsilon_k^q)}{D_{\max}^q \ln \left[\frac{T_f \ln(1/\varepsilon_k^q)}{\lambda_k D_{\max}^q} + 1 \right]} \text{ (packets/s)}. \quad (\text{A.2})$$

IPP: The effective bandwidth of the IPP can be expressed as [45]

$$E_k^B(\theta_k) = \frac{\Omega}{2\theta_k T_f} \text{ (packet/s)}, \quad (\text{A.3})$$

where $\Omega \triangleq [(e^{\theta_k} - 1) \lambda_k^{\text{on}} - (\alpha + \beta)] + \sqrt{[(e^{\theta_k} - 1) \lambda_k^{\text{on}} - (\alpha + \beta)]^2 + 4\alpha (e^{\theta_k} - 1) \lambda_k^{\text{on}}}$. Substituting (A.3) into (11), the QoS exponent θ_k can be obtained from $\Omega = \frac{-2T_f \ln \varepsilon_k^q}{D_{\max}^q}$ numerically.

SPP: Deriving the effective bandwidth of autocorrelated processes is much harder than that of renewal processes. To overcome this difficulty, we provide an upper bound of the effective bandwidth of SPP. Without loss of generality, we assume $\lambda_k^{\text{I}} \leq \lambda_k^{\text{II}}$.

Consider a Poisson process with average arrival rate λ_k^{II} , the arrival rate in the first state of SPP is less than that of the Poisson process. Thus, the effective bandwidth of the SPP is less than that of the Poisson process, which can be obtained by substituting $\lambda_k = \lambda_k^{\text{II}}$ into (A.1).

APPENDIX B

UPPER BOUND OF THE PACKET DROPPING PROBABILITY

Proof. To derive ε_k^h , we introduce an upper bound of $b_k^d(n)$ as follows,

$$b_k^U(n) = \begin{cases} \max(T_f E_k^B(\theta_k) - s_k^{\text{th}}, 0), & \text{if } Q_k(n) > 0, \\ 0, & \text{if } Q_k(n) = 0, \end{cases}$$

considering that $b_k^U(n) = b_k^d(n)$ when $Q_k(n) \geq T_f E_k^B(\theta_k)$ or $Q_k(n) = 0$, and $b_k^U(n) > b_k^d(n)$ when $0 < Q_k(n) < T_f E_k^B(\theta_k)$. Then, we can derive an upper bound of $\mathbb{E}[b_k^d(n)]$ as

$$\mathbb{E}[b_k^U(n)] = \eta_k \int_0^{\frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}} (T_f E_k^B(\theta_k) - s_k^{\text{th}}) f_g(g) dg.$$

Substituting $\mathbb{E}[b_k^U(n)]$ into (21), we obtain an upper bound of the packet dropping probability as

$$\varepsilon_k^h \leq \int_0^{\frac{N_0 W_k \gamma_k}{\mu_k P_k^{\text{th}}}} \left[1 - \frac{s_k^{\text{th}}}{T_f E_k^B(\theta_k)} \right] f_g(g) dg, \quad (\text{B.1})$$

where $\eta_k = \Pr\{Q_k(n) > 0\} = \mathbb{E}\{\sum_{i \in \mathcal{A}_k} a_i(n)\} / \mathbb{E}[s_k(n)] = \mathbb{E}\{\sum_{i \in \mathcal{A}_k} a_i(n)\} / [T_f E_k^B(\theta_k)]$ is applied.

By substituting s_k^{th} in (19) and considering (17), we have

$$\frac{s_k^{\text{th}}}{T_f E_k^B(\theta_k)} \approx \frac{\ln\left(1 + \frac{\mu_k P_k^{\text{th}} g_k}{N_0 W_k}\right) - \sqrt{\frac{V}{\phi W_k}} f_Q^{-1}(\varepsilon_k^c)}{\ln(1 + \gamma_k) - \sqrt{\frac{V}{\phi W_k}} f_Q^{-1}(\varepsilon_k^c)}. \quad (\text{B.2})$$

Because a packet is dropped only if it will be transmitted in deep fading, i.e. $g_k \rightarrow 0$, V in (4) approaches 0, and then (B.2) can be further accurately approximated by

$$\frac{s_k^{\text{th}}}{T_f E_k^B(\theta_k)} \approx \frac{\ln\left(1 + \frac{\mu_k P_k^{\text{th}} g_k}{N_0 W_k}\right)}{\ln(1 + \gamma_k)}. \quad (\text{B.3})$$

Substituting (B.3) into (B.1), we obtain the upper bound in (22). \square

APPENDIX C

PROOF OF THE CONVEXITY OF THE OBJECTIVE FUNCTION IN (24)

Proof. For the Q-function $f_Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{\tau^2}{2}\right) d\tau$, we have $f_Q'(x) \triangleq -\frac{1}{\sqrt{2\pi}} e^{-x^2/2} < 0$, and $f_Q''(x) = \frac{x}{\sqrt{2\pi}} e^{-x^2/2} > 0$ when $x > 0$. Thus, $f_Q(x)$ is an decreasing and strictly convex function when $x > 0$, i.e. $f_Q(x) < 0.5$. Since the inverse function of a decreasing and strictly convex function is also strictly convex [43], $f_Q^{-1}(\varepsilon^c)$ is strictly convex when $\varepsilon^c < 0.5$ (which is true for any application). Hence, the second term of (24) is strictly convex.

To prove that the first term of (24) is strictly convex, we first derive its second order derivative. Denote $y = -\ln(\varepsilon^q)$ and $z = \frac{T_f}{D_{\max}^q \lambda} > 0$. After removing the non-relevant constants, the first term of (24) can be expressed as $f(y) = \frac{y}{\ln(1+zy)}$. The second order derivative of $f(y)$ can be derived as

$$\frac{d^2 f}{d(\varepsilon^q)^2} = \left(\frac{d^2 f}{dy^2} \right) \left(\frac{dy}{d\varepsilon^q} \right)^2 + \left(\frac{df}{dy} \right) \left(\frac{d^2 y}{d(\varepsilon^q)^2} \right). \quad (\text{C.1})$$

After some regular derivations, we can obtain that

$$\frac{dy}{d\varepsilon^q} = -\frac{1}{\varepsilon^q}, \quad \frac{d^2 y}{d(\varepsilon^q)^2} = \left(\frac{1}{\varepsilon^q} \right)^2, \quad (\text{C.2})$$

$$\frac{df}{dy} = \frac{(1+zy) \ln(1+zy) - zy}{[\ln(1+zy)]^2 (1+zy)}, \quad \frac{d^2 f}{dy^2} = \frac{2z^2 y - (2z + z^2 y) \ln(1+zy)}{[\ln(1+zy)]^3 (1+zy)^2}. \quad (\text{C.3})$$

After substituting (C.2) and (C.3) into (C.1), we can finally obtain that

$$\frac{d^2 f}{d(\varepsilon^q)^2} = \frac{(1+zy)^2 [\ln(1+zy)]^2 - (2z + zy + z^2 y + z^2 y^2) \ln(1+zy) + 2z^2 y}{[\ln(1+zy)]^3 (1+zy)^2 (\varepsilon^q)^2}. \quad (\text{C.4})$$

Since the denominator is positive, we only need to show the numerator is positive. Denote the numerator of (C.4) as $f_{\text{mun}}(x, z)$, where $x = yz$. Then, we have

$$f_{\text{mun}}(x, z) = (1+x)^2 [\ln(1+x)]^2 - (x+x^2) \ln(1+x) - [(2+x) \ln(1+x) - 2x] z. \quad (\text{C.5})$$

For $\varepsilon^q < 10^{-5}$, which is true for applications with ultra-high reliability requirement, $y > -\ln(10^{-5}) > 10$, and then $x > 10z$. Moreover, $(2+x) \ln(1+x) - 2x > 0, \forall x > 0$. Then, we can obtain a lower bound of $f_{\text{mun}}(x, z)$ as follows,

$$f_{\text{LB}}(x) = (1+x)^2 [\ln(1+x)]^2 - (x+x^2) \ln(1+x) - [(2+x) \ln(1+x) - 2x] x/10. \quad (\text{C.6})$$

When $x = 0$, $f_{\text{LB}}(x) = 0$. To prove $f_{\text{LB}}(x) > 0, \forall x > 0$, we substitute $\nu = x + 1$ into (C.6) and prove $f'_{\text{LB}}(\nu) > 0, \forall \nu > 1$. It is not hard to derive that

$$f'_{\text{LB}}(\nu) = \frac{20\nu^2(\ln \nu)^2 + (10\nu - 2\nu^2) \ln \nu + (3\nu - 11)(\nu - 1)}{10\nu}. \quad (\text{C.7})$$

Denote the numerator of (C.7) as $f_{\text{LBnum}}(\nu)$, which equals zero when $\nu = 0$. Besides,

$$f'_{\text{LBnum}}(\nu) = 40\nu(\ln \nu)^2 + (10 + 36\nu) \ln \nu + 4(\nu - 1) > 0, \forall \nu > 1.$$

As a result, $f'_{\text{LB}}(\nu) > 0$, and hence $f_{\text{LB}}(x)$ increases with x . Therefore, we have $f_{\text{LB}}(x) > 0, \forall x > 0$. This completes the proof. \square

APPENDIX D

PROOF OF THE OPTIMALITY OF THE TWO-STEP METHOD

Proof. Denote an arbitrary feasible solution of problem (23) and the related transmit power as $(\tilde{\varepsilon}^a, \tilde{\varepsilon}^c, \tilde{\varepsilon}^h)$ and \tilde{P}^{\max} , respectively. Given $\tilde{\varepsilon}^h$, we can obtain the global minimal transmit power $P^{\max}(\tilde{\varepsilon}^h) \leq \tilde{P}^{\max}$ by solving problem (24), which is for Poisson arrival process. In the second step, the global optimal ε^{h*} is obtained such that $P^{\max*} \leq P^{\max}(\tilde{\varepsilon}^h)$. Therefore, $P^{\max*} \leq \tilde{P}^{\max}$. The proof follows. \square

APPENDIX E

ACHIEVABLE RATE OVER FREQUENCY-SELECTIVE CHANNEL

Denote the channel vector on the j th subchannel of the k th user as $\mathbf{h}_{kj} \in \mathbb{C}^{N_t \times 1}$. Then, the channel matrix over frequency-selective channel can be equivalent to a $N_t N_k^s \times N_k^s$ MIMO channel with bandwidth W_c , i.e.,

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{h}_{k1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{h}_{k2} & \dots & \mathbf{0} \\ & & \dots & \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{h}_{kN_k^s} \end{bmatrix}, \quad (\text{E.1})$$

and $\mathbf{H}_k^H \mathbf{H}_k = \text{diag}(g_{k1}, g_{k2}, \dots, g_{kN_k(n)})$, where $g_{kj} = \mathbf{h}_{kj}^H \mathbf{h}_{kj}$ is the channel gain on the j th subchannel allocated to the k th user and also one of the eigenvalues of $\mathbf{H}_k^H \mathbf{H}_k$. Then, by substituting the eigenvalues into (96) and (97) in [27], the number of packets that can be transmitted in one frame can be expressed as (26).

REFERENCES

- [1] C. She, C. Yang, and T. Quek, "Cross-layer transmission design for tactile internet," in *Proc. IEEE Global Commun. Conf. (Globecom)*, 2016.
- [2] 3GPP, *Study on Scenarios and Requirements for Next Generation Access Technologies*. Technical Specification Group Radio Access Network, Technical Report 38.913, Release 14, Oct. 2016.
- [3] G. P. Fettweis, "The tactile internet: Applications & challenges," *IEEE Vehic. Tech. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

- [4] P. Popovski, *et al.*, “Deliverable d6.3 intermediate system evaluation results.” ICT-317669-METIS/D6.3, 2014. [Online]. Available: https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D6.3_v1.pdf
- [5] 3GPP, *Further Advancements for E-UTRA Physical Layer Aspects*. Technical Specification Group Radio Access Network, Technical Report 36.814, Release 9, Mar. 2010.
- [6] A. Osseiran, F. Boccardi and V. Braun, *et al.*, “Scenarios for 5G mobile and wireless communications: The vision of the METIS project,” *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May. 2014.
- [7] S.-Y. Lien, S.-C. Hung, K.-C. Chen, and Y.-C. Liang, “Ultra-low-latency ubiquitous connections in heterogeneous cloud radio access networks,” *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 22–31, Jun. 2015.
- [8] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, “Downlink packet scheduling in LTE cellular networks: Key design issues and a survey,” *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2013.
- [9] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, “5G-enabled tactile internet,” *IEEE J. Select. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.
- [10] S. A. Ashraf, F. Lindqvist, R. Baldemair, and B. Lindoff, “Control channel design trade-offs for ultra-reliable and low-latency communication system,” in *IEEE Global Commun. Conf. (Globecom) Workshops*, 2015.
- [11] P. Kela and J. Turkka, *et al.*, “A novel radio frame structure for 5G dense outdoor radio access networks,” in *Proc. IEEE Veh. Tech. Conf. (VTC) Spring*, 2015.
- [12] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
- [13] K. Niu, K. Chen, J. Lin, and Q. T. Zhang, “Polar codes: Primary concepts and practical decoding algorithms,” *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 192–203, Jul. 2014.
- [14] D. Ohmann, M. Simsek, and G. P. Fettweis, “Achieving high availability in wireless networks by an optimal number of Rayleigh-fading links,” in *IEEE Global Commun. Conf. (Globecom) Workshops*, 2014.
- [15] F. Kirsten, D. Ohmann, M. Simsek, and G. P. Fettweis, “On the utility of macro- and microdiversity for achieving high availability in wireless networks,” in *Symp. IEEE Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, 2015.
- [16] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen, and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks,” in *IEEE Global Commun. Conf. (Globecom) Workshops*, 2015.
- [17] O. N. C. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, “Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case,” in *IEEE Int. Conf. on Commun. (ICC) Workshops*, 2015.
- [18] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, “Radio access for ultra-reliable and low-latency 5G communications,” in *IEEE Int. Conf. on Commun. (ICC) Workshops*, 2015.
- [19] C. Chang and J. A. Thomas, “Effective bandwidth in high-speed digital networks,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.
- [20] D. Wu and R. Negi, “Effective capacity: A wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
- [21] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks,” in *IEEE Global Commun. Conf. (Globecom) Workshops*, Dec. 2014.
- [22] A. Aijaz, “Towards 5G-enabled tactile internet: Radio resource allocation for haptic communications,” in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2016.
- [23] S. Schiessl, J. Gross, and H. Al-Zubaidy, “Delay analysis for wireless fading channels with finite blocklength channel coding,” in *Proc. ACM MSWiM*, 2015.

- [24] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, 2011.
- [25] S. Xu, T.-H. Chang, S.-C. Lin, C. Shen, and G. Zhu, "On the convexity of energy-efficient packet scheduling problem with finite blocklength codes," in *IEEE Global Commun. Conf. (Globecom) Workshops*, 2015.
- [26] R. A. Berry, "Optimal power-delay tradeoffs in fading channels—small-delay asymptotics," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3939–3952, Jun. 2013.
- [27] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4264, Jul. 2014.
- [28] W. Whitt, "Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues," *Telecommunication Systems*, vol. 2, no. 1, pp. 71–107, 1993.
- [29] J. Wu, Y. Bao, G. Miao, S. Zhou, and Z. Niu, "Base station sleeping control and power matching for energy-delay tradeoffs with bursty traffic," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3657–3675, May 2016.
- [30] G. Zhang, T. Q. S. Quek, M. Kountouris, A. Huang, and H. Shan, "Fundamentals of heterogeneous backhaul design—analysis and optimization," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 876–889, Feb. 2016.
- [31] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 49–55, Apr. 2014.
- [32] C. She, C. Yang, and T. Q. S. Quek, "Uplink transmission design with massive machine type devices in tactile internet," in *IEEE Globecom Workshops*, 2016.
- [33] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.
- [34] M. Khabazian, S. Aissa, and M. Mehmet-Ali, "Performance modeling of safety messages broadcast in vehicular ad hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 380–387, Mar. 2013.
- [35] G. R1-120056, "Analysis on traffic model and characteristics for MTC and text proposal." Technical Report, TSG-RAN Meeting WG1#68, Dresden, Germany, 2012.
- [36] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.
- [37] H. A. Omar, W. Zhuang, A. Abdrabou, and L. Li, "Performance evaluation of VeMAC supporting safety applications in vehicular networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 1, no. 1, pp. 69–83, Aug. 2013.
- [38] D. Leonardis, M. Solazzi, I. Bortone, and A. Frisoli, "A 3-RSR haptic wearable device for rendering fingertip contact forces," *IEEE Trans. Haptics*, early access.
- [39] D. Gross and C. Harris, *Fundamentals of Queueing Theory*. Wiley, 1985.
- [40] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 203–217, Feb. 1996.
- [41] I. E. Telatar, *Capacity of multi-antenna Gaussian channels*, 1995.
- [42] I.-H. Hou, V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," in *Proc. IEEE Int. Conf. on Comput. Commun. (INFOCOM)*, 2009.
- [43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
- [44] B. Soret, M. C. Aguayo-Torres, and J. T. Entrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated Rayleigh channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 1901–1911, Jun. 2010.
- [45] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1375–1395, Mar. 2016.